Electronic Journal of Statistics
Vol. 3 (2009) 176–204
ISSN: 1935-7524
DOI: 10.1214/08-EJS254

General oracle inequalities for model selection

Charles Mitchell

e-mail: mitchell@stat.math.ethz.ch

Sara van de Geer

e-mail: geer@stat.math.ethz.ch

Abstract: Model selection is often performed by empirical risk minimization. The quality of selection in a given situation can be assessed by risk bounds, which require assumptions both on the margin and the tails of the losses used. Starting with examples from the 3 basic estimation problems, regression, classification and density estimation, we formulate risk bounds for empirical risk minimization and prove them at a very general level, for general margin and power tail behavior of the excess losses. These bounds we then apply to typical examples.

AMS 2000 subject classifications: Primary 62G05; secondary 62G20.

Received June 2008.

1. Introduction

Consider a sample $Z_1, \ldots, Z_N$ of independent random variables in some space $\mathcal{Z}$, whose distribution depends on an unknown parameter f. To estimate f, we split the sample into two parts: a test set $Z_1, \ldots, Z_n$ and a training set $Z_{n+1}, \ldots, Z_N$. Based on the training set various estimators of f are constructed, say $\hat{f}_1, \ldots, \hat{f}_p$. To decide among these estimators, we use the test set. Suppose that $\gamma_f : \mathcal{Z} \to \mathbf{R}$ is a loss function. The final estimate $\hat{f}$ is now chosen to minimize the empirical risk:

$$\hat{f} := \arg \min_{\hat{f}_j : 1 \le j \le p} \frac{1}{n} \sum_{i=1}^{n} \gamma_{\hat{f}_j}(Z_i) \; .$$

In this paper, we examine whether this empirical risk minimization leads to taking, among the p estimators, the “nearly best” one. Here, “nearly best” will be defined in terms of the excess risk of the estimators.

The behavior of the excess risk near f will be called the margin behavior. We not only consider the classical case, which is quadratic margin behavior, but also more general margin behavior. For the tails of our excess loss functions, we consider both an exponential moment condition and a more general power tail condition. We prove a risk inequality under the most general combination of these conditions, and in doing so automatically obtain risk inequalities for more restricted situations. These latter situations represent examples we give from regression, classification and density estimation.

A common and succinct way of expressing the quality of an aggregated estimator is by way of an oracle inequality of the form

$$\mathbf{E}R(\hat{\gamma}) \leq A \cdot \inf_{\gamma \in \boldsymbol{\Gamma}} R(\gamma) + C(\boldsymbol{\Gamma}, n) \ .$$

Here $R(\gamma) := \mathbf{E}_Z \gamma(Z)$ is the risk of the procedure that has loss γ, and $C(\boldsymbol{\Gamma}, n)$ is a quantity that depends on the cardinality (when finite) or complexity (such as the metric entropy) of the class $\boldsymbol{\Gamma}$ of models or aggregates up for selection, as well as on the sample size n.

When the number of procedures being aggregated is a finite number $p := |\boldsymbol{\Gamma}|$, most of the results in the literature set $O(\log(p)/n)$ to be the benchmark for the rate of the term $C(\boldsymbol{\Gamma}, n)$ above. For instance, Bunea et al. [8] give this rate for Gaussian regression and a linear aggregate that minimizes a penalized sum of squares. For a more general risk problem, Györfi and Wegkamp [11] obtain a similar result, and Lecué [15] achieves the same rate for the Cumulative Aggregation with Exponential Weights (CAEW) procedure in a classification setup with bounded loss. Other types of results in this vein include Bartlett and Mendelson's [5] high probability bounds for the estimator risk of empirical risk minimization, done for the estimation of functions from a class with a uniform bound.

The analysis of empirical risk minimization stands on two major pillars. The first of these is empirical process theory. In Vapnik and Chervonenkis' seminal work on pattern recognition [25], the importance of the empirical process

$$((P_n - P)(f))_{f \in \mathcal{F}}$$

of the class $\mathcal{F}$ of candidate procedures for the study of empirical risk minimizers was already recognized. More recently, van de Geer [22] also describes the use of empirical processes in understanding such estimators. The second foundation we need is the study of concentration inequalities, which describe the concentration of random variables and their empirical means around their true means. The value of such inequalities in the analysis of model selection via empirical risk minimization is recognised, and put to use, in the papers of Barron et. al. [4] and Birgé and Massart [6].

In much of the literature, the quantities to be estimated are assumed to be uniformly bounded. Another very important condition for ensuring good rates in oracle inequalities is the margin condition, which controls the "noise" between procedures that differ only very slightly in risk, and thus makes assumptions on the *small-scale* behaviour of the family of losses. For some regression setups, a uniform bound on the target and the estimates already dispenses with the need for a margin condition, as in the results of Bunea et al. [8]. (We shall see in Example 3.1 that such a uniform bound implies the margin condition when using L^2-loss.) In classification, though, which is the original area for margin conditions, the situation is somewhat more complex. Here the margin conditions that hold are generally weaker than the ones known in regression or density estimation setups. Tsybakov [21] provides a good treatment of this case.

Koltchinskii [14] looks at a wider range of situations, generalizing Tsybakov's results, among others; besides a margin condition, his approach also requires direct conditions on the empirical process or on the complexity of the candidate class Γ in lieu of boundedness conditions. In this paper, we shall define the margin condition in Section 3 and there examine it more closely.

Generally, most of the literature deals with only one particular problem, such as regression; furthermore, the strong boundedness conditions usually imposed are not always necessary. It is well-known that some conditions must be imposed in order to obtain risk rates that are better than $O(1/\sqrt{n})$. For example, Lee et al. [17] give an overview of risk rates in an agnostic learning setup and show that convexity properties on the class of candidate functions lead to risk rates around $O(1/n)$ rather than $O(1/\sqrt{n})$. Mendelson [19] uses a least-squares regression example to also show that $O(1/\sqrt{n})$ cannot be improved upon without assuming something like a Bernstein-type inequality. (While convexity assumptions can suffice for obtaining fast risk rates, they are not always necessary, as also shown by Mendelson [18]). Our interest lies in inequalities for a general loss function setup, with boundedness conditions replaced by suitably loose requirements on the tails, at least when conditioning on the training set. Such conditioning on the training set is common practice; to average the results over the training data then requires margin and power tail conditions to hold uniformly over all trained versions of the estimators used, if possible – or if not, then other, possibly more stringent conditions.

Another fairly general approach is taken by Audibert [1], who looks at the general prediction problem, i.e. regression and classification, and uses a progressive mixture rule for aggregation, but with only a brief reference to averaging over the training stage, which would be part of the full sample splitting problem. On the other hand, Rigollet [20] examines sample splitting schemes with multiple splits and thus comes close to cross validation, but does so only for the problem of density estimation. A direct treatment of a cross validation scheme is to be found in van der Vaart et al. [24]. And in the context of classification, recent inequalities are given for recursive aggregation by mirror descent by Juditsky et al. [13] and for aggregation with exponential weights by Lecué [15].

1.1. Notation

The results will be conditional on the training set. We use $\mathbf{P}$ to denote the distribution of the test sample, and $\mathbf{E}$ denotes the expectation of random variables depending on the test sample.

For $\gamma : \mathcal{Z} \to \mathbf{R}$, we write

$$P\gamma := \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\gamma(\tilde{Z}_i) \ ,$$

where $(\tilde{Z}_1, \ldots, \tilde{Z}_n)$ is an i.i.d. copy of $(Z_1, \ldots, Z_n)$, and

$$P_n \gamma := \frac{1}{n} \sum_{i=1}^{n} \gamma(Z_i) \ .$$

Let $\gamma_j : \mathcal{Z} \to \mathbf{R}$, $j = 1, \ldots, p$, be given loss functions in a class $\mathbf{\Gamma}$. Given the training set, γ_j may be taken as short-hand (and slight abuse of) notation for $\gamma_{\hat{f}_j}$, $j = 1, \ldots, p$. We consider the empirical risk minimization estimator

$$\hat{\gamma} := \arg \min_{1 \le j \le p} P_n \gamma_j \ .$$

The target is

$$\gamma_0 := \arg \min_{\gamma \in \mathbf{\Gamma}} P\gamma \ ,$$

whose best approximation is

$$\gamma_* := \arg \min_{1 \le j \le p} P\gamma_j \ .$$

We will write f_* for the corresponding parameter value (or an arbitrary choice thereof, if it is not unique) at which this minimum is attained, i.e. for which $\gamma_* = \gamma_{f_*}$. We define the excess risks

$$\hat{\mathcal{E}} := P(\hat{\gamma} - \gamma_0)$$

(which is a random variable, as it depends on the test sample),

$$\mathcal{E}_j := P(\gamma_j - \gamma_0)$$

and

$$\mathcal{E}_* := P(\gamma_* - \gamma_0) \ .$$

Without loss of generality, we assume that $\mathbf{\Gamma}$ is of the form $\mathbf{\Gamma} := \{\gamma_f : \ f \in \mathbf{F}\}$, where $\mathbf{F}$ is a subset of a semi-metric space with semi-metric d, and write (with some abuse of notation) γ_{f_j} as γ_j, $\{f_j\}_{j=1}^p \subset \mathbf{F}$.

1.2. Goal

Our goal is now to show that $\hat{\mathcal{E}}$ is close to $\mathcal{E}_*$ (with large probability or in expectation). The results are modifications of inequalities of the form

$$(1 - \delta)\mathbf{E}\hat{\mathcal{E}} \le (1 + \delta)\mathcal{E}_* + \frac{\Delta}{\delta} \ ,$$

where $\delta > 0$ is an arbitrary small constant, and with Δ of order $\log(p)/n$ and not depending on $\mathcal{E}_*$ – see for example Chapter 7 in Györfi et al. [10]. In the standard setup of Section 4 and under a quadratic "margin condition", for instance, we show that for $1 \le m \le 1 + \log p$

$$\mathbf{E}\hat{\mathcal{E}}^{\frac{m}{2}} \le \left(\sqrt{\mathcal{E}_*} + \sqrt{\Delta}\right)^m \ ,$$

with Δ of order $\log(2p)/n$ and not depending on $\mathcal{E}_*$. In particular, with $m = 2$, this reads

$$\mathbf{E}\hat{\mathcal{E}} \le \left(\sqrt{\mathcal{E}_*} + \sqrt{\Delta}\right)^2 \ .$$

This gives rise to a non-sharp oracle inequality

$$\mathbf{E}\hat{\mathcal{E}} \le (1+\delta)\mathcal{E}_* + \Delta \ , \quad \delta > 0 \ .$$

A sharp ($\delta = 0$) and rate-optimal (correction term $O(\Delta)$) oracle inequality cannot be established in a general setup by empirical risk minimization (cf. Lecué [15]). Instead, methods such as mirror averaging could be used, as by Juditsky et al. [12]. See also Audibert ([2] and [3]) for some limitations of empirical risk minimization, and alternative approaches to overcome the limitations. We however believe empirical risk minimization remains an important topic of study because it is widely applied in practice, and is closely related to various cross validation schemes.

1.3. Convex loss

In our proofs, we only use the property

$$P_n\hat{\gamma} \le P_n\gamma_* \ .$$

In the convex case, this sometimes means that conditions can be weakened. Let $\mathbf{F}$ be a convex subset of a linear vector space, and suppose that $\mathbf{\Gamma} := \{\gamma_f : \ f \in \mathbf{F}\}$, with $f \mapsto \gamma_f$ convex, $\mathbf{P}$-almost everywhere. Then for $0 \le \alpha \le 1$, we have the inequality

$$P_n\gamma_{\alpha\hat{f}+(1-\alpha)f_*} \le \alpha P_n\hat{\gamma} + (1-\alpha)P_n\gamma_* \le P_n\gamma_* \ .$$

This means that we can replace $\hat{\gamma}$ by $\gamma_{\alpha\hat{f}+(1-\alpha)f_*}$ throughout, leading to inequalities for the excess risk

$$\hat{\mathcal{E}}_\alpha = P\gamma_{\alpha\hat{f}+(1-\alpha)f_*} - P\gamma_0 \ .$$

From these, one may then often deduce inequalities for the original $d(\hat{f}, f_0)$. As we shall see, this extension (with $\alpha < 1$) allows us to work with weaker conditions (than with $\alpha = 1$). In particular, the example on maximum likelihood will take a similar approach with α set to $1/2$.

1.4. Organization of the paper

The paper is organized as follows. Section 2 presents Bernstein's inequality. It is stated in the form of a probability inequality and a moment inequality. Section 3 presents the margin condition and some examples where it holds. Section 4 gives the main results, both one for exponential moments and a very general margin condition, and one for power tails and a particular form of margin condition. Subsequently, Section 5 applies the main results to the examples already given. Finally, the proofs are in Section 6.

2. Bernstein's inequality

Bernstein's inequality for a single average is well known, and the extension of Bernstein's probability inequality to a uniform probability inequality over p averages is completely straightforward. The result can be seen as the simplest version of a concentration inequality in the spirit e.g. of Bousquet [7] (emphasizing how tight these general concentration inequalities are). The moment inequality for the maximum of p averages is perhaps less known.

For all j, we let

$$\gamma_j^c(\cdot) := \gamma_j(\cdot) - P\gamma_j$$

denote the centered loss functions. To obtain our results, we make assumptions on the tails of the centered excess losses $\gamma_j^c - \gamma_*^c$ or of their envelope $\Gamma := \max_{1 \le j \le p} \left|\gamma_j^c - \gamma_*^c\right|$ as follows:

Definition 2.1. *We say that the excess losses $\gamma_j - \gamma_*$ satisfy the exponential moment condition for some $K > 0$ if*

$$P\left|\gamma_j^c - \gamma_*^c\right|^m \le \frac{m!}{2}(2K)^{m-2} d^2(f_j, f_*) \tag{1}$$

for all $m = 2, 3, \dots$ and for all $j = 1, \dots, p$.

We say that the envelope function Γ has power tails of order $s > 1$ if there exists an $M \in (0, \infty)$ such that

$$\mathbf{P}\left(\Gamma > K\right) \le \left(\frac{M}{K}\right)^s \quad \forall K > 0\ . \tag{2}$$

Here $d(\cdot,\cdot)$ is a semi-metric on the underlying parameter space that allows for different weighting of the procedures under consideration. As an important example of this define, for all γ, the variance

$$\sigma^2(\gamma) := P|\gamma^c|^2\ .$$

Then clearly (1) implies that

$$d^2(f_j, f_*) \ge \sigma^2(\gamma_j - \gamma_*)\ \forall\ j\ . \tag{3}$$

Moreover, if the bound $|\gamma_j - \gamma_*| \le 3K$ holds for all j, then (1) holds with

$$d^2(f_j, f_*) = \sigma^2(\gamma_j - \gamma_*)\ \forall\ j.$$

In the following sections, we will indeed often assume (1) with this value for $d(f_j, f_*)$, but we will also consider an extension. The choice of the semi-metric d is intertwined with the margin behavior, which we consider in the next section. Furthermore, when applying the margin condition, we shall implicitly use the inequality (3). As we will make repeated use of Bernstein's inequality, and the term $2\log(2p)/n$ will appear frequently, we will henceforth denote this term by

$$\Delta := \frac{2\log(2p)}{n}\ .$$

Using this notation, the version of Bernstein's inequality that we will need in this paper is:

Lemma 2.1. *(Bernstein's inequality for the maximum of p averages: weighted version) Assume that for some constant K, the exponential moment condition (1) holds. Then for all $t > 0$ and $\tau > 0$*

$$\mathbf{P}\left(\max_{1\le j\le p}\frac{|P_n(\gamma_j^c-\gamma_*^c)|}{d(f_j,f_*)\vee\tau}\ge\sqrt{\Delta+2t/n}+\frac{K(\Delta+2t/n)}{\tau}\right)\le\exp[-t]\ .$$

Moreover, for all $1\le m\le 1+\log p$,

$$\left(\mathbf{E}\max_{1\le j\le p}\left(\frac{|P_n(\gamma_j^c-\gamma_*^c)|}{d(f_j,f_*)\vee\tau}\right)^m\right)^{1/m}\le\sqrt{\Delta}+\frac{K\Delta}{\tau}\ .$$

Remark: The moment inequality is for moments of order $m\le 1+\log p$. It can be extended to hold for general m, provided a slight adjustment, depending on m, is made on the constants. Because we have the situation in mind where p is large, we have formulated the result for $m\le 1+\log p$ to facilitate the exposition.

3. Margin behavior

Definition 3.1. *We say that the margin condition holds with strictly convex margin function $G(\cdot)$ if*

$$P(\gamma_j-\gamma_0)\ge G\left(d(f_j,f_0)\right),\ \forall\ j\ . \tag{4}$$

Furthermore, we say that the margin condition holds with constants $\kappa > 1/2$ and $C > 0$, if (4) holds with

$$G(u)=u^{2\kappa}/C^{2\kappa},\ \ u>0\ .$$

The specific case of $G(u)=u^{2\kappa}/C^{2\kappa}$ is the one most typically used in the literature, with the semi-metric d taken to be the variance of the excess loss $\gamma_j-\gamma_0$. Such a margin condition can be found e.g. in Chesneau and Lecué [9] for regression and density estimation setups, with a comparable result as here for regression, and a result for a different example (squared loss) given for density estimation. Tsybakov [21] gives a similar margin condition for classification. In that paper, the use of 0-1-loss means that $d(f_j,f_0)=P_X(G_{f_j}\triangle G_{f_0})$, where G_f denotes the set $\{x : f(x)=1\}$. The concept of a Bernstein class – as used by Bartlett and Mendelson [5] – is the same thing after a suitable reparametrization. As we shall see, $\kappa=1$ in typical cases – but other, in particular larger, values can also occur.

Let us now consider some examples. In a regression or classification situation, we have i.i.d. random pairs $Z_i=(X_i,Y_i)$, with $Y_i\in\mathcal{Y}\subset\mathbf{R}$ a response variable, and $X_i\in\mathcal{X}$ a covariable, $i=1,\ldots,n$. The quality of an estimator f of $\mathbf{E}[Y_i|X_i]$ can be measured by applying a loss function $\gamma:\mathcal{Y}\times\mathcal{Y}\to\mathbf{R}$ to the true and the estimated response.

Example 3.1. *(Regression)* Suppose that $\{Z_i\}_{i=1}^n := \{(X_i, Y_i)\}_{i=1}^n$. Let $\mathbf{F}$ be a class of real-valued functions on $\mathcal{X}$, and for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, let

$$\gamma_f(x, y) := \gamma(f(x), y), \ \ f \in \mathbf{F} \ .$$

Set

$$l(a, \cdot) = \mathbf{E}(\gamma(a, Y_i)|X_i = \cdot), \ \ a \in \mathbf{R} \ .$$

We moreover write $l_f(x) := l(f(x), x)$. As our target we take the overall minimizer

$$f_0(\cdot) := \arg\min_{a \in \mathbf{R}} l(a, \cdot) \ .$$

We now check whether the margin condition holds with $\kappa = 1$ and

$$d^2(f, f_0) := K_2^2 P|f - f_0|^2 \ ,$$

where K_2 is an appropriate constant.

Lemma 3.1. *Assume that for some $K_1 > 0$, and all $|f - f_0| \leq K_1$,*

$$l_f - l_{f_0} \geq (f - f_0)^2 / C_0^2 \ . \tag{5}$$

Then

$$P(\gamma_f - \gamma_{f_0}) \geq d^2(f, f_0)/C^2 \ ,$$

where $C^2 := C_0^2 K_2^2$. If we moreover assume (for $i = 1, \ldots, n$) that

$$\text{var}(\gamma_f(Z_i) - \gamma_{f_0}(Z_i)) \leq K_2^2 \mathbf{E}(f(X_i) - f_0(X_i))^2 \ , \tag{6}$$

then for all $\|f - f_0\|_\infty \leq K_1$, we have

$$\sigma^2(\gamma_f - \gamma_{f_0}) \leq d^2(f, f_0) \ .$$

If $l(a, \cdot)$ has two derivatives near $a = f_0(\cdot)$, and the second derivatives are positive and bounded away from zero, then $l(a, \cdot)$ behaves quadratically near its minimum, i.e., then (5) holds for some $K_1 > 0$.

It also also clear that (6) holds as soon as $\gamma(\cdot, y)$ is Lipschitz for all y, with Lipschitz constant L. Then we may take $K_2 = L$. When $\gamma(\cdot, y)$ is not Lipschitz (e.g., quadratic loss), it may be useful to define

$$e_f(Z_i) := \gamma(f(X_i), Y_i) - l_f(X_i) \ .$$

Then obviously

$$\text{var}(\gamma_f(Z_i) - \gamma_{f_0}(Z_i)) = \text{var}(e_f(Z_i) - e_{f_0}(Z_i)) + \text{var}(l_f(X_i) - l_{f_0}(X_i)) \ . \tag{7}$$

Note that with fixed design, the second term in (7) vanishes.

Quadratic loss:
In the case of least squares, the loss function is

$$\gamma(f, y) := (y - f)^2 \ ,$$

Then

$$l_f - l_{f_0} = |f - f_0|^2 \ ,$$

and

$$e_f(Z_i) - e_{f_0}(Z_i) = 2\epsilon_i(f(X_i) - f_0(X_i)) \ ,$$

with $\epsilon_i := Y_i - f_0(X_i)$. Assuming that the conditional variance is bounded by some constant σ_ϵ, i.e.,

$$\max_{1\le i\le n} \text{var}(Y_i|X_i) \le \sigma_\epsilon^2 \ , \tag{8}$$

we may conclude the following.

Least squares with fixed design:
The margin condition holds with $\kappa = 1$ and $C^2 = 4\sigma_\epsilon^2$.

Least squares with random design:
If $\|f_j - f_0\|_\infty \le K_1$ for all j, the margin condition holds with $\kappa = 1$ and $C^2 = 4\sigma_\epsilon^2 + K_1^2$.

Example 3.2. *(Classification)* Suppose that $Z_i = (X_i, Y_i)$, with $Y_i \in \mathcal{Y} := \{0,1\}$ a label, $i = 1, \ldots, n$. Let $\mathbf{F}$ be a class of functions $f : \mathcal{X} \to \{0,1\}$. We consider 0/1-loss

$$\gamma_f(x,y) = \gamma(f(x),y) := (1-y)f(x) + y(1-f(x)), \ f \in \mathbf{F}, \ (x,y) \in \mathcal{X} \times \{0,1\} \ .$$

For $a \in [0,1]$, write

$$\begin{aligned} l(a,\cdot) &:= \mathbf{E}(\gamma(a,Y_i)|X_i = \cdot) \\ &= (1-\eta)a + \eta(1-a) = a(1-2\eta) + \eta \ , \end{aligned}$$

where $\eta = \mathbf{E}(Y_i|X_i = \cdot)$. The target is again the overall minimizer

$$f_0 := \arg\min_{a\in\{0,1\}} l(a,\cdot) \ .$$

It is clear that f_0 is the Bayes rule

$$f_0 = \text{l}\{1 - 2\eta < 0\} \ .$$

We moreover have

$$P(\gamma_f - \gamma_{f_0}) = P|(f - f_0)(1 - 2\eta)| \ .$$

Consider the function

$$H_1(v) \le vP\text{l}\{|1-2\eta| < v\}, \ v \in [0,1]$$

and its convex conjugate

$$G_1(u) = \max_v \{uv - H_1(v)\}, \ u \in [0,1]$$

(assuming the maximum exists).

Lemma 3.2. *The inequality*

$$P(\gamma_f - \gamma_{f_0}) \geq G\big(\sigma(\gamma_f - \gamma_{f_0})\big)$$

holds with $G(u) = G_1(u^2)$, $u \in [0,1]$.

If $H_1(v) = 0$ for $v \leq C_1$, we take $G_1(u) = C_1 u$. More generally, the Tsybakov margin condition (see [21]) assumes that one may take, for some $C_1 \geq 1$ and $\lambda \geq 0$ (Tsybakov himself writes γ for this parameter),

$$H_1(v) = v(C_1 v)^{1/\lambda} \ ,$$

Then one has

$$G_1(u) = u^{1+\lambda}/C^{1+\lambda} \ ,$$

where

$$C = C_1^{\frac{1}{1+\lambda}} \lambda^{-\frac{\lambda}{1+\lambda}} (1+\lambda) \ .$$

Thus, then the margin condition holds with this value of C and with $\kappa = 1 + \lambda$ (for $d(f_j, f_0) = \sigma(\lambda_j - \lambda_0)$, $\forall \ j$).

Example 3.3. *(Maximum likelihood)* Suppose that $\{Z_i\}_{i=1}^n$ are iid. with density $f_0 := dP/d\mu$, where μ is a σ-finite dominating measure. Let $\mathbf{F}$ be a convex class of densities w.r.t. μ, containing f_0. Consider the transformed log-likelihood loss

$$\gamma_f(\cdot) := \gamma(f(\cdot)),$$

where $\gamma(a) = -\log(a)/2$. Define

$$\bar{f} = (f + f_*)/2, \ f \in \mathbf{F} \ .$$

The squared Hellinger distance of densities f and $\tilde{f}$ is

$$h^2(f, \tilde{f}) = \frac{1}{2} \int \left(\sqrt{f} - \sqrt{\tilde{f}} \right)^2 d\mu, \ f, \tilde{f} \in \mathbf{F} \ .$$

We now check the margin and power tail conditions for a distance measure $d(f, f_0)$ which is a multiple of $h(f, f_0)$.

Lemma 3.3. *For all densities f, we have*

$$P(\gamma_f - \gamma_{f_0}) \geq h^2(f, f_0) \ .$$

Moreover, under the assumption

$$\sqrt{\frac{f_0}{f_*}} \leq \frac{L}{4} \ ,$$

we have

$$P|\gamma_{\bar{f}} - \gamma_{f_*}|^m \leq \frac{m!}{2} L^2 h^2(\bar{f}, f_*) \ .$$

This lemma contains the exponential moment condition (1) for $K = 1$, and also allows us to deduce the margin condition

$$\sigma(\gamma_{\bar{f}} - \gamma_{f_*}) \leq \left[P(\gamma_{\bar{f}} - \gamma_{f_*})^2 \right]^{1/2} \leq L h(\bar{f}, f_*) \ .$$

for margin constants $\kappa = 1$ and $C = 1$.

4. Main results

If we assume exponential tails on the loss functions, we are able to obtain a result for a wide range of margin conditions:

Lemma 4.1. *Let G be a strictly convex and increasing function with G(0)=0. Suppose that the margin condition holds for G. Let H be the convex conjugate of G, i.e.*

$$H(v) = \sup_{u \geq 0} [uv - G(u)] \quad \forall v \geq 0 \ .$$

Assume that for some $m \leq 1 + \log p$, the function $H(v^{\frac{1}{m}})$, $v > 0$, is concave. Assume moreover that the exponential moment condition (1) holds for some $K > 0$ and for $d(f_j, f_*) := G^{-1}(\mathcal{E}_j) + G^{-1}(\mathcal{E}_*)$. Then for all $0 < \delta < 1$, and $\varepsilon > 0$, we have

$$(1-\delta)\mathbf{E}\hat{\mathcal{E}} \leq 2\delta H\left(\frac{\sqrt{\Delta}}{\delta} + \frac{K\Delta}{2\delta G^{-1}(\mathcal{E}_* \vee \varepsilon)}\right) + (1+\delta)(\mathcal{E}_* \vee \varepsilon) \ .$$

The next theorem focuses on the common family of margin functions $G(u) = u^{2\kappa}/C^{2\kappa}, u > 0, \kappa \geq 1$, but also relaxes the exponential tail condition to a power tail condition. Note that for this family of margin functions, the corresponding convex conjugates are $H(v)$ of order $O(v^{\frac{2\kappa}{2\kappa-1}})$, and thus Lemma 4.1 gives an oracle inequality with correction term rate $O(\Delta^{\frac{\kappa}{2\kappa-1}})$, which agrees with the rates found in the literature and in the next theorem:

Theorem 4.1. *(i) Suppose that the margin condition holds for the loss functions γ_j with constants $\kappa \geq 1$ and $C > 0$ and some d satisfying $d(f_j, f_0) \geq \sigma(\gamma_j - \gamma_0)$, $\forall\ j$. Also assume that the envelope Γ has power tails in the form of (2), of order $s > 1$ and for some $M > 0$. Then for all m in the interval $[2\kappa, \min(2s\kappa, 1 + \log(p))[$ and for all $\tau > 0$, we have the following inequality:*

$$\begin{aligned}\left\| (\hat{\mathcal{E}})^{\frac{1}{2\kappa}} \right\|_m \leq\ & (\mathcal{E}_* \vee \tau)^{\frac{1}{2\kappa}} + A(\kappa) \cdot C^{\alpha} \cdot \Delta^{\alpha/2} \\ & + \xi(\kappa, s, m) \cdot M^{\frac{s}{m} \cdot \frac{\alpha}{\alpha+\beta}} \cdot \Delta^{\frac{\alpha\beta}{\alpha+\beta}} \cdot (\mathcal{E}_* \vee \tau)^{-\frac{1}{2\kappa} \cdot \frac{\alpha\beta}{\alpha+\beta}} \quad ,\end{aligned}$$

where

$$\alpha := \frac{1}{2\kappa - 1} \ , \quad \beta := \frac{s}{m} - \frac{1}{2\kappa} \ ,$$

$$A(\kappa) := \frac{1 + (2\kappa-1)^{\frac{1}{2\kappa-1}}}{\kappa^{\frac{1}{2\kappa-1}}}$$

and

$$\xi(\kappa, s, m) := A(\kappa)^{\frac{\beta}{\alpha+\beta}} \cdot 2^{\frac{1}{2\kappa} \cdot \frac{\alpha}{\alpha+\beta}} \cdot \left(\frac{2s\kappa}{2s\kappa - m}\right)^{\frac{\alpha}{\alpha+\beta} \cdot \frac{1}{m}} \cdot \left(\left(\frac{\beta}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}} + \left(\frac{\alpha}{\beta}\right)^{\frac{\beta}{\alpha+\beta}}\right).$$

(ii) *Furthermore, if the excess losses satisfy the exponential moment condition (1) for some constant $K > 0$, then*

$$\left\| (\hat{\mathcal{E}})^{\frac{1}{2\kappa}} \right\|_m \leq (\mathcal{E}_* \vee \tau)^{\frac{1}{2\kappa}} + A(\kappa) \cdot \left(C \cdot \sqrt{\Delta} + \frac{K\Delta}{(\mathcal{E}_* \vee \tau)^{\frac{1}{2\kappa}}} \right)^{\alpha}$$

for all m in the interval $[2\kappa, 1 + \log(p)[$. In this case we also have tail bounds

$$\mathbf{P}\left(\hat{\mathcal{E}}^{\frac{1}{2\kappa}} \geq (\mathcal{E}_* \vee \tau)^{\frac{1}{2\kappa}} + A(\kappa) \left(C\sqrt{\Delta + 2t/n} + \frac{K(\Delta + 2t/n)}{(\mathcal{E}_* \vee \tau)^{\frac{1}{2\kappa}}} \right)^{\alpha} \right) \leq \mathrm{e}^{-t}$$

for all $t > 0$.

These statements lead to simpler ones if we use that $\tau \leq \mathcal{E}_* \vee \tau \leq \mathcal{E}_* + \tau$ and then optimize over τ, trading off the summand with positive exponent $1/(2\kappa)$ and the one with negative exponent $-1/(2\kappa) \cdot \alpha\beta/(\alpha + \beta)$. This yields the main result of this paper:

Corollary 4.1. *Under the conditions of Theorem 4.1, we have the inequalities*

$$(i) \qquad \left\| (\hat{\mathcal{E}})^{\frac{1}{2\kappa}} \right\|_m \leq \mathcal{E}_*^{\frac{1}{2\kappa}} + A(\kappa) \cdot C^{\alpha} \cdot \Delta^{\alpha/2} + \tilde{\xi}(\kappa, s, m) \cdot M^{\frac{s}{m} \cdot \frac{\alpha}{\alpha+\beta+\alpha\beta}} \cdot \Delta^{\frac{\alpha\beta}{\alpha+\beta+\alpha\beta}}$$

when the loss envelope Γ has power tails (2) ($\tilde{\xi}(\kappa, s, m)$ is a constant depending only on κ, s and m), and

$$(ii) \qquad \left\| (\hat{\mathcal{E}})^{\frac{1}{2\kappa}} \right\|_m \leq \mathcal{E}_*^{\frac{1}{2\kappa}} + A(\kappa) \cdot C^{\alpha} \cdot \Delta^{\alpha/2} + \left(A(\kappa)^{2\kappa - 1} \cdot K\Delta \right)^{\frac{1}{2\kappa}}$$

when the excess losses satisfy the exponential moment condition (1). In the latter case we also have the tail bound

$$\mathbf{P}\Big(\hat{\mathcal{E}}^{\frac{1}{2\kappa}} \geq \mathcal{E}_*^{\frac{1}{2\kappa}} + A(\kappa) \cdot C^{\alpha} \cdot (\Delta + 2t/n)^{\alpha/2} + \left(A(\kappa)^{2\kappa-1} \cdot K(\Delta + 2t/n) \right)^{\frac{1}{2\kappa}} \Big) \leq \mathrm{e}^{-t}$$

for all $t > 0$.

Note: These risk inequalities yield oracle inequalities in quite a natural manner: In general, if we have an inequality

$$\left\| (\hat{\mathcal{E}})^{\frac{1}{2\kappa}} \right\|_m \leq \mathcal{E}_*^{\frac{1}{2\kappa}} + \xi$$

that holds for a range of values of m including $m = 2\kappa$, then this latter choice of m gives a further inequality

$$\left(\mathbf{E}\hat{\mathcal{E}} \right)^{\frac{1}{2\kappa}} \leq \mathcal{E}_*^{\frac{1}{2\kappa}} + \xi \ ,$$

and for any $\delta > 0$ the general inequality $(a + b)^{2\kappa} \leq (1 + \delta) \cdot a^{2\kappa} + (1 + 1/\delta) \cdot b^{2\kappa}$ (for $a, b \geq 0$ and $\delta > 0$) then yields the oracle inequality

$$\mathbf{E}\hat{\mathcal{E}} \leq (1 + \delta) \cdot \mathcal{E}_* + \left(1 + \frac{1}{\delta} \right) \cdot \xi^{2\kappa} \ .$$

Corollary 4.1 naturally also leads to statements about risk ratios. Under the exponential moment condition, for example, we can see that when

$$\mathcal{E}_* \gg \max\{A(\kappa)^{2\kappa} \cdot C^{\frac{2\kappa}{2\kappa-1}} \cdot \Delta^{\frac{\kappa}{2\kappa-1}}, A(\kappa)^{2\kappa-1} \cdot K\Delta\} ,$$

we have have the ratio inequality

$$\mathbf{E}\left|\frac{\hat{\mathcal{E}}}{\mathcal{E}_*}\right|^{\frac{m}{2\kappa}} \to 1$$

for all $m \in [1, 1+\log(p)]$.

The results of Corollary 4.1 constitute a generalization of other, similar, results to be found in the literature. For instance, the rate $O(\Delta^{\frac{\kappa}{2\kappa-1}})$ we obtain for exponential tails and a margin condition of order $\kappa \geq 1$ is similar to that described by Lecué [16] for classification using Tsybakov's margin condition; the only difference is that there the rate also depends on that of the oracle, i.e. the rate at which $\mathcal{E}_*$ tends to zero as Δ does. For bounded losses, Chesneau and Lecué [9] give a general oracle inequality that they subsequently apply to examples of density estimation and bounded regression. Their most general oracle inequality also has the rate $O(\Delta^{\frac{\kappa}{2\kappa-1}})$ when the oracle rate is not too large.

5. Application to examples

We can apply Corollary 4.1 to the (more restricted) cases described in the previous sections:

5.1. Quadratic margin, exponential tails

The quadratic margin condition corresponds to taking $\kappa = 1$. Taking the second part of Corollary 4.1 for this value of κ yields the oracle inquality

$$P\hat{\mathcal{E}} \leq (1+\delta)\mathcal{E}_* + \left(1+\frac{1}{\delta}\right) \cdot 2\left(C+\sqrt{K}\right)^2 \cdot \Delta \tag{9}$$

for all $\delta > 0$, when the losses satisfy the exponential moment condition.

Example 5.1. *(Maximum Likelihood)*

Take the setup of Example 3.3 and assume that

$$\sqrt{\frac{f_0}{f_*}} \leq \frac{L}{4} . \tag{10}$$

Define new parameters $\bar{f}_j = (f_j + f_*)/2$ and Kullback-Leibler information numbers

$$\hat{\mathcal{K}} := P(\gamma_{(\hat{f}+f_*)/2} - \gamma_{f_0}) = \hat{\mathcal{E}}_{1/2}$$

and

$$\mathcal{K}_* := P(\gamma_{f_*} - \gamma_{f_0}) = \mathcal{E}_* \ .$$

In Lemma 3.3, we have already shown the margin and exponential moment conditions for the transformed parameters $\bar{f}_j$ and the scaled Hellinger distance $d(f, f') := Lh(f, f')$. The parameters there are $C = K = \kappa = 1$, and thus we obtain the oracle inequality

$$\mathbf{E}\hat{\mathcal{K}} \leq (1+\delta)\mathcal{K}_* + 8\left(1 + \frac{1}{\delta}\right) \cdot \Delta \ . \tag{11}$$

This involves the density $(\hat{f} + f_*)/2$, which is *not* an estimator. We can however use this oracle inequality to deduce a risk inequality for the estimator $\hat{f}$ using the following lemma about the Hellinger distance:

Lemma 5.1. *Let f,f' and f_0 be densities with respect to the measure μ. Then we have the following inequality:*

$$h(f, f_0) \leq (2+\sqrt{2})h\left(\frac{f+f'}{2}, f_0\right) + (1+\sqrt{2})h(f', f_0) \ .$$

By the first part of Lemma 3.3, we have $\hat{\mathcal{K}} \geq h^2(\bar{f}, f_0)$ and $\mathcal{K}_* \geq h^2(f_*, f_0)$. Combining this with the oracle inequality (11) and with Lemma 5.1, we obtain the risk inequality

$$\begin{aligned}
\mathbf{E}h^2(f, f_0) \quad &\leq \quad 2(2+\sqrt{2})^2\mathbf{E}h^2(\bar{f}, f_0) + 2(1+\sqrt{2})^2 h^2(f_*, f_0) \\
&\leq \quad 2(2+\sqrt{2})^2\mathbf{E}\hat{\mathcal{K}} + 2(1+\sqrt{2})^2\mathcal{K}_* \\
&\leq \quad \left[2(2+\sqrt{2})^2(1+\delta) + 2(1+\sqrt{2})^2\right] \cdot \mathcal{K}_* \\
&\quad\quad +16(2+\sqrt{2})^2\left(1 + \frac{1}{\delta}\right) \cdot \Delta \ .
\end{aligned}$$

We cannot expect to obtain an oracle inequality involving $\mathcal{E}_*$, however, as there is no general bound of the Kullback-Leibler distance of densities by their Hellinger distance.

5.2. *Quadratic margin, power tails*

Here $\kappa = 1$ and thence $\alpha = 1$, $\beta = s/m - 1/2$ and $A(\kappa) = 2$. Corollary 4.1 thus implies

$$\left\|\sqrt{\hat{\mathcal{E}}}\right\|_m \leq \sqrt{\mathcal{E}_*} + 2C \cdot \sqrt{\Delta} + \tilde{\xi}(1, s, m) \cdot \sqrt{M} \cdot \Delta^{\frac{1}{2} - \frac{m}{4s}} \ ,$$

and for $m = 2$ and any $\delta > 0$, the oracle inequality

$$\mathbf{E}\hat{\mathcal{E}} \leq (1+\delta)\mathcal{E}_* + \left(1 + \frac{1}{\delta}\right) \cdot 4\left(C^2 \cdot \Delta + \tilde{\xi}(1, s, 2)^2 \cdot M \cdot \Delta^{1-\frac{1}{s}}\right) \tag{12}$$

holds.

Example 5.2. *(Regression)*

Upper bounds:

In Example 3.1, we saw that least-squares regression satisfies a quadratic margin condition, i.e. one with $\kappa = 1$. For instance, we have the margin parameter $C := 2\sigma_\epsilon$ in the fixed-design case. If furthermore we assume that the errors ϵ_i possess some finite moment of order $2s > 2$ – a less restrictive assumption than the Gaussianity often required – then the loss has power tails of order $s > 1$:

$$\gamma_f(x, y) = \gamma(f(x), y) = (y - f(x))^2 = (\epsilon + f_0(x) - f(x))^2$$

$$\Rightarrow \mathbf{E}[\Gamma^s] \le 2^s \mathbf{E}\Big[\sup_{f\in F} \left|\gamma_f^c(X, Y)\right|^s\Big] \le 2^{4s-1} \cdot \mathbf{E}\Big[|\epsilon|^{2s} + \sup_{f\in F} |f_0(X) - f(X)|^{2s}\Big]$$

$$= 2^{4s-1} \cdot \Big(\mathbf{E}|\epsilon|^{2s} + \mathbf{E}\sup_{f\in F} |f_0(X) - f(X)|^{2s}\Big) =: M \ ,$$

and so by Chebyshev,

$$P(\{\Gamma > K\}) \le \frac{\mathbf{E}[\Gamma^s]}{K^s} \le \left(\frac{M}{K}\right)^s \quad \forall K > 0 \ .$$

Thus the oracle inequality (12) holds here.

Lower bounds:

Consider the fixed-design case with double Pareto tails of order $s > 2$, i.e. the distribution of the ϵ_i is symmetric around 0, and

$$P(|\epsilon_i| \le u) = 1 - \frac{1}{(1+u)^s}, \ u > 0 \ .$$

Fix some $p \in \mathbf{N}$, $p \ge 2$, and define $f_p := f_0 \equiv 0$,

$$f_j(x) = \mathrm{l}\{x = X_j\} n^{\frac{1}{2s}}, \ x \in \mathcal{X}, \ j = 1, \ldots, p-1 \ .$$

Thus $f_* \equiv 0$ and $\mathcal{E}_* = 0$, too.

Lemma 5.2. *The margin condition holds with $\kappa = 1$ and $C^2 = 8/((s-2)(s-1))$, and when $p \ge \sqrt{n}+1$, the power tail condition (2) holds with $M = 2$. For $n \ge 2^{2s}$ and all $p \le n$, moreover, we have*

$$\hat{\mathcal{E}} \ge n^{-\frac{s-1}{s}}$$

with probability at least $1 - \exp[-2^{-1} \cdot (p-1)/\sqrt{n}]$.

Remark: We can easily extend the lower bound result to $p > n$, because we can add, as candidates, any number of bounded functions f_j, say with $\|f_j\|_\infty \le 1$, without neccessitating an increase in the scale parameter M of the power tail condition. These additional functions may be selected by the least squares estimator, but if they all have norm $Pf_j^2 \geq n^{-\frac{s-1}{s}}$, selecting one of these still gives the same lower bound.

Combining this lower bound with the oracle inequality (12), we find that for $n \geq p \geq \sqrt{n}+1$, we have

$$n^{-\frac{s-1}{s}} \leq \mathbf{E}\hat{\mathcal{E}} \leq C'(s) \cdot \left(\frac{\log(n)}{n}\right)^{-\frac{s-1}{s}}$$

for some constant $C'(s)$ that depends only on s, which shows the rate-optimality – up to a logarithmic factor – of the upper bound. If p is small compared to $\sqrt{n}$, however, things look very different:

Lemma 5.3. *We have*

$$\left\| \sqrt{\hat{\mathcal{E}}} \right\|_s \leq \sqrt{\mathcal{E}_*} + Cc_s p^{1/s} M/\sqrt{n} \ ,$$

where

$$c_s := 2\sqrt{\frac{2}{\pi}}\Gamma^{1/s}\left(\frac{s+1}{2}\right) \ .$$

This leads to a (non-sharp) oracle inequality whose correction term has the order $p^{2/s}/n$. If $p \ll \sqrt{n}$, then $p^{2/s}/n \ll n^{-\frac{s-1}{s}}$, i.e. a lower bound of order $n^{-\frac{s-1}{s}}$ for $\mathbf{E}\hat{\mathcal{E}}$ will not hold.

5.3. General margin, exponential tails

The risk bound in this case was given in Part (ii) of Corollary 4.1, whose correction term is of order $O(\Delta^{1/(4\kappa-2)})$. Taking $m = 2\kappa$, this leads to an oracle inequality of the form

$$\mathbf{E}\hat{\mathcal{E}} \leq (1+\delta)\mathcal{E}_* + \left(1+\frac{1}{\delta}\right) O\left(\Delta^{\frac{\kappa}{2\kappa-1}}\right)$$

for all $\delta > 0$.

Example 5.3. *(Classification)*
In Example 3.2, we saw the margin condition for $\kappa = 1+\lambda$ and

$$C = C_1^{\frac{1}{1+\lambda}} \lambda^{-\frac{\lambda}{1+\lambda}} (1+\lambda) \ ,$$

where $\lambda \geq 0$, as a consequence of Tsybakov's margin condition. Furthermore,

$$\begin{aligned} P\left|\gamma_f^c - \gamma_{f_0}^c\right|^m &= P\left|(f(X)-f_0(X)) \cdot (1-2Y) - P\left|(f-f_0)(1-2\eta)\right|\right|^m \\ &\leq 2^{m-2} \cdot P\left|\gamma_f^c - \gamma_{f_0}^c\right|^2 = 2^{m-2} \cdot \sigma^2\left(\gamma_f - \gamma_{f_0}\right) \end{aligned}$$

for all f in this example, which means that the excess losses have exponential moments (1) with $K = 1$. Thus we have an oracle inequality

$$\begin{aligned}\mathbf{E}\hat{\mathcal{E}} \quad &\leq \quad (1+\delta)\mathcal{E}_* + \left(1+\frac{1}{\delta}\right)\left(\tilde{A}_1(C_1,\lambda)\cdot\Delta^{\frac{1+\lambda}{1+2\lambda}} + \tilde{A}_2\cdot\Delta\right)\\ &= (1+\delta)\mathcal{E}_* + \left(1+\frac{1}{\delta}\right)O(\Delta^{\frac{1+\lambda}{1+2\lambda}})\end{aligned}$$

for all $\delta > 0$, and for constants $\tilde{A}_1(C,\lambda)$ and $\tilde{A}_2$.

6. Proofs

6.1. Proofs for Section 2

Proof of Lemma 2.1. Without loss of generality, suppose that $\mathbf{E}\gamma_j(Z_i) = 0$ for all i and j. Furthermore we can reduce to the case where $\gamma_* \equiv 0$ and all $d(f_j, f_*) = 1$ by looking at a new set of loss functions $(\gamma_j - \gamma_*)/(d(f_j, f_*) \vee \tau)$, where $\tau > 0$ is arbitrary. Thus it suffices to show that under the condition that

$$P|\gamma_j|^m \leq \frac{m!}{2}(2K)^{m-2},\ m = 2, 3, , \ldots$$

for centered loss functions γ_j, the inequality

$$\mathbf{P}\left(\max_{1\leq j\leq p}|P_n\gamma_j| \geq \sqrt{\frac{2(\log(2p)+t)}{n}} + \frac{2K(\log(2p)+t)}{n}\right) \leq \exp[-t] \tag{13}$$

holds for all $t > 0$, and that for all $1 \leq m \leq 1 + \log p$,

$$\left(\mathbf{E}\left(\max_{1\leq j\leq p}|P_n\gamma_j|\right)^m\right)^{1/m} \leq \sqrt{\frac{2\log(2p)}{n}} + \frac{2K\log(2p)}{n}\ . \tag{14}$$

Bernstein's probability inequality says that for all $t > 0$,

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n \gamma_j(Z_i) \geq 2Kt + \sqrt{2t}\right) \leq \exp\left[-nt\right], \forall\ j\ . \tag{15}$$

This inequality follows from the intermediate result

$$\mathbf{E}\exp\left[\sum_{i=1}^n \gamma_j(Z_i)/L\right] \leq \exp\left[\frac{n}{2(L^2-2LK)}\right],\ \forall\ j\ , \tag{16}$$

which holds for all $L > 2K$. Inequality (13) follows immediately from (15).

To prove (14), we apply Lemma 6.1 to the function $g : x \mapsto (L\cdot\log(x+1))^m$, which is increasing on $[0,\infty)$ and concave on $[\mathrm{e}^{m-1}-1,\infty)$. We then obtain for all $L > 0$ and all m that

$$\mathbf{E}\left(\max_j\left|\sum_{i=1}^n \gamma_j(Z_i)\right|^m\right) \leq L^m\log^m\left[\mathbf{E}\exp\left[\max_j\left|\sum_{i=1}^n\gamma_j(Z_i)\right|/L - 1\right] + \mathrm{e}^{m-1} - 1\right].$$

From (16), and invoking $e^{|x|} \leq e^{x} + e^{-x}$, we obtain for $L > 2K$,

$$L^m \log^m \left[\mathbf{E} \exp \left[\max_j \left| \sum_{i=1}^n \gamma_j(Z_i) \right| / L - 1 \right] + e^{m-1} - 1 \right]$$

$$\leq L^m \log^m \left[p \left\{ 2 \exp \left[\frac{n}{2(L^2 - 2LK)} \right] - 1 \right\} + e^{m-1} \right]$$

$$\leq L^m \log^m \left[(2p + e^{m-1} - p) \exp \left[\frac{n}{2(L^2 - 2LK)} \right] \right]$$

$$= \left(L \log(2p + e^{m-1} - p) + \left[\frac{n}{2(L - 2K)} \right] \right)^m .$$

Now take

$$L = 2K + \sqrt{\frac{n}{2 \log(2p + e^{m-1} - p)}}$$

and use the extra restriction $m \leq 1 + \log p$ to get the desired result. □

Lemma 6.1. *(Jensen's inequality for partly concave functions) Let X be a real-valued random variable, and let g be an increasing function on $[0, \infty)$, which is concave on $[c, \infty)$ for some $c \geq 0$. Then*

$$\mathbf{E} g(|X|) \leq g\big[\mathbf{E}|X| + c\mathbf{P}(|X| < c)\big] .$$

Proof. We have

$$\mathbf{E} g(|X|) = \mathbf{E} g(|X|) \mathrm{l}\{|X| \geq c\} + \mathbf{E} g(|X|) \mathrm{l}\{|X| < c\}$$

$$\leq \mathbf{E} g(|X|) \mathrm{l}\{|X| \geq c\} + g(c)\mathbf{P}(|X| < c)$$

$$= \mathbf{E} \left[g(|X|) \Big| |X| \geq c \right] \mathbf{P}(|X| \geq c) + g(c)\mathbf{P}(|X| < c) .$$

We now apply Jensen's inequality to the term on the left, and then use the concavity on $[c, \infty)$ to incorporate the term on the right:

$$\mathbf{E} g(|X|) \leq g \left[\mathbf{E} \left(|X| \Big| |X| \geq c \right) \right] \mathbf{P}(|X| \geq c) + g(c)\mathbf{P}(|X| < c)$$

$$\leq g\big[\mathbf{E}|X| + c\mathbf{P}(|X| < c)\big] .$$

□

6.2. *Proofs for Section 3*

Proof of Lemma 3.1. This follows from

$$P(\gamma_f - \gamma_{f_0}) = P(l_f - l_{f_0}) .$$

□

Proof of Lemma 3.2. We have

$$\begin{aligned} P|(f-f_0)(1-2\eta)| &\geq vP|f-f_0|\mathrm{l}\{|1-2\eta|\geq v\} \\ &\geq v\left(P|f-f_0| - P\mathrm{l}\{|1-2\eta|<v\}\right) := uv - H_1(v) \ , \end{aligned}$$

with $u = P|f-f_0|$. Since this is true for all v, we may maximize over v to obtain

$$P|(f-f_0)(1-2\eta)| \geq G_1\bigl(P|f-f_0|\bigr) \geq G_1\bigl(P(f-f_0)^2\bigr) \ ,$$

as

$$P|f-f_0| \geq P(f-f_0)^2 \ .$$

Moreover,

$$|\gamma_f(y) - \gamma_{f_0}(y)| = |(f-f_0)(1-2y)| \leq |f-f_0| \ ,$$

so that

$$\sigma^2(\gamma_f - \gamma_{f_0}) \leq P(\gamma_f - \gamma_{f_0})^2 \leq P(f-f_0)^2 \ .$$

□

Proof of Lemma 3.3. As the excess risk is a Kullback-Leibler distance to the true distribution, the first statement of the Lemma is just the classical lower bound by the Hellinger distance:

$$\begin{aligned} P(\gamma_f - \gamma_{f_0}) &= -\int_{f_0>0} \log\sqrt{\frac{f}{f_0}} f_0 d\mu \\ &\geq -\int_{f_0>0}\left(\sqrt{\frac{f}{f_0}}-1\right) f_0 d\mu \\ &= 1 - \int \sqrt{ff_0}d\mu = h^2(f,f_0) \ . \end{aligned}$$

For the second part, we can use Lemma 7.2 in van de Geer [22], which says that

$$\exp|\gamma_{\bar f} - \gamma_{f_*}| - |\gamma_{\bar f} - \gamma_{f_*}| - 1 \leq 4\left(\sqrt{\frac{\bar f}{f_*}}-1\right)^2 . \tag{17}$$

We moreover have

$$|\gamma_{\bar f} - \gamma_{f_*}|^m \leq \frac{m!}{2}\{\exp|\gamma_{\bar f} - \gamma_{f_*}| - |\gamma_{\bar f} - \gamma_{f_*}| - 1\} \ .$$

Thus

$$P|\gamma_{\bar f} - \gamma_{f_*}|^m \leq 2m!\int(\sqrt{\bar f} - \sqrt{f_*})^2\frac{f_0}{f_*}d\mu \leq \frac{m!}{2}L^2 h^2(\bar f, f_*) \ .$$

□

6.3. Proofs for Section 4

6.3.1. Preparatory lemmas

We begin with two simple results (without proofs) for ease of reference.

Lemma 6.2. *If the loss envelope Γ has power tails (2), then for all $m < 2s$ and $K > 0$,*

$$P\Gamma^{m/2}\mathrm{l}\{\Gamma > K\} \leq \frac{2s}{2s-m} M^s K^{-(2s-m)/2} \ .$$

Lemma 6.3. *For positive constants a, b, α and β, the function*

$$g(x) := ax^{\alpha} + bx^{-\beta},\ x > 0$$

is minimized at

$$x_0 := \left(\frac{b\beta}{a\alpha}\right)^{\frac{1}{\alpha+\beta}} \ ,$$

and there attains a minimum of

$$g(x_0) = \tilde{\mathcal{C}}(\alpha, \beta) \cdot a^{\frac{\beta}{\alpha+\beta}} b^{\frac{\alpha}{\alpha+\beta}} \ ,$$

where

$$\tilde{\mathcal{C}}(\alpha, \beta) := \left(\frac{\beta}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}} + \left(\frac{\alpha}{\beta}\right)^{\frac{\beta}{\alpha+\beta}} \ .$$

Next we need a couple of auxiliary lemmas:

Lemma 6.4. *For all $0 \leq z \leq 1$, we have the inequality*

$$(1-z)^{2\kappa} \leq 1 - 2\kappa z^{2\kappa-1} + (2\kappa-1)z^{2\kappa} \ ,$$

and for all $z \geq 0$,

$$(1+z)^{2\kappa} \geq 1 + 2\kappa z^{2\kappa-1} + z^{2\kappa} \ .$$

Proof. The second part is clear, as it involves the omission only of positive summands from the LHS to the RHS. For the first part, we write

$$f(z) := 1 - 2\kappa z^{2\kappa-1} + (2\kappa-1)\, z^{2\kappa} - (1-z)^{2\kappa}$$

and note that

$$\begin{aligned}
f(z) &= 1 - z^{2\kappa} - (1-z) \cdot 2\kappa z^{2\kappa-1} - (1-z)^{2\kappa} \\
&= (1-z) \cdot \left(\sum_{i=0}^{2\kappa-1} z^j - 2\kappa z^{2\kappa-1} - (1-z)^{2\kappa-1}\right) \\
&= (1-z)^2 \cdot \left(\sum_{j=0}^{2\kappa-2} (j+1)\, z^j - (1-z)^{2\kappa-2}\right) \\
&=: (1-z)^2 \cdot \tilde{f}(z) \ .
\end{aligned}$$

Now as $\tilde{f}(0) = 0$ and for $0 \leq z \leq 1$,

$$\left(\tilde{f}\right)'(z) = \sum_{j=1}^{2\kappa-2} j(j+1)z^{j-1} + (2\kappa - 2) \cdot (1-z)^{2\kappa-3} \geq 0 \ ,$$

we know that $\tilde{f}(z)$, and thus $f(z)$, is non-negative on $[0, 1]$. □

Lemma 6.5. *Let a, b and c be positive, let $\kappa \geq 1$, and assume that*

$$a \leq b + c \cdot \left(a^{\frac{1}{2\kappa}} + b^{\frac{1}{2\kappa}}\right) \ .$$

Then

$$a^{\frac{1}{2\kappa}} \leq \left(1 + (2\kappa - 1)^{\frac{1}{2\kappa-1}}\right) \cdot \left(\frac{c}{2\kappa}\right)^{\frac{1}{2\kappa-1}} + b^{\frac{1}{2\kappa}} \ .$$

Proof. First note that if $a^{1/2\kappa} \leq (c/2\kappa)^{1/(2\kappa-1)}$, then the desired inequality automatically holds. Thus we can restrict ourselves to the case where $a^{1/2\kappa} > (c/2\kappa)^{1/(2\kappa-1)}$. Applying the first part of Lemma 6.4 for $z = (c/2\kappa)^{1/(2\kappa-1)}/a^{1/2\kappa}$ – which now is less than 1 – gives us the inequality

$$\left(a^{\frac{1}{2\kappa}} - \left(\frac{c}{2\kappa}\right)^{\frac{1}{2\kappa-1}}\right)^{2\kappa} - \left(\frac{1}{2\kappa-1}\right)\left(\frac{c}{2\kappa}\right)^{\frac{2\kappa}{2\kappa-1}} \leq a - c \cdot a^{\frac{1}{2\kappa}} \leq b + c \cdot b^{\frac{1}{2\kappa}} \ ,$$

and thus

$$\begin{aligned}\left(a^{\frac{1}{2\kappa}} - \left(\frac{c}{2\kappa}\right)^{\frac{1}{2\kappa-1}}\right)^{2\kappa} &\leq b + c \cdot b^{\frac{1}{2\kappa}} + (2\kappa - 1)\left(\frac{c}{2\kappa}\right)^{\frac{2\kappa}{2\kappa-1}} \\ &\leq b + (2\kappa-1)\, c \cdot b^{\frac{1}{2\kappa}} + \left(\frac{2\kappa-1}{2\kappa} \cdot c\right)^{\frac{2\kappa}{2\kappa-1}} \ ,\end{aligned}$$

where in the second step we used that $\kappa \geq 1$. Now part 2 of Lemma 6.4, applied to $z = \left(\frac{2\kappa-1}{2\kappa} \cdot c\right)^{1/2\kappa-1}/b^{1/2\kappa}$, yields

$$\begin{aligned}\left(b^{\frac{1}{2\kappa}} + \left(\frac{2\kappa-1}{2\kappa} \cdot c\right)^{\frac{1}{2\kappa-1}}\right)^{2\kappa} &\geq b + (2\kappa-1)\, c \cdot b^{\frac{1}{2\kappa}} + \left(\frac{2\kappa-1}{2\kappa} \cdot c\right)^{\frac{2\kappa}{2\kappa-1}} \\ &\geq \left(a^{\frac{1}{2\kappa}} - \left(\frac{c}{2\kappa}\right)^{\frac{1}{2\kappa-1}}\right)^{2\kappa} \ ,\end{aligned}$$

from which the stated inequality follows. □

6.3.2. Main proofs

Proof of Lemma 4.1. Define

$$\mathbf{Z} := \frac{|(P_n - P)(\hat{\gamma} - \gamma_*)|}{G^{-1}(\hat{\mathcal{E}}) + G^{-1}(\mathcal{E}_* \vee \varepsilon)} \ .$$

By the definition of the convex conjugate H, we have

$$H\left(\frac{\mathbf{Z}}{\delta}\right) \geq \mathbf{Z} \cdot G^{-1}(\hat{\mathcal{E}}) - \delta \cdot \hat{\mathcal{E}}$$

and

$$H\left(\frac{\mathbf{Z}}{\delta}\right) \geq \mathbf{Z} \cdot G^{-1}(\mathcal{E}_* \vee \varepsilon) - \delta \cdot (\mathcal{E}_* \vee \varepsilon) \ .$$

Then

$$\begin{aligned}\hat{\mathcal{E}} &\leq \mathbf{Z}G^{-1}(\hat{\mathcal{E}}) + \mathbf{Z}G^{-1}(\mathcal{E}_* \vee \varepsilon) + \mathcal{E}_* \\ &\leq \delta\hat{\mathcal{E}} + 2\delta H\left(\frac{\mathbf{Z}}{\delta}\right) + (1+\delta)(\mathcal{E}_* \vee \varepsilon) \ .\end{aligned}$$

It follows that

$$\begin{aligned}(1-\delta)\mathbf{E}\hat{\mathcal{E}} &\leq 2\delta\mathbf{E}H\left(\frac{\mathbf{Z}}{\delta}\right) + (1+\delta)(\mathcal{E}_* \vee \varepsilon) \\ &\leq 2\delta H\left(\mathbf{E}\left(\frac{\mathbf{Z}}{\delta}\right)^m\right)^{1/m} + (1+\delta)(\mathcal{E}_* \vee \varepsilon) \ .\end{aligned}$$

Now as $d(f_j, f_*) = G^{-1}(\mathcal{E}_j) + G^{-1}(\mathcal{E}_*)$, we have the upper bound

$$P\left|\frac{(\gamma_j - \gamma_*)}{G^{-1}(\mathcal{E}_j) + G^{-1}(\mathcal{E}_* \vee \varepsilon)}\right|^{\tilde{m}} \leq \frac{\tilde{m}!}{2}\left(2\frac{K}{G^{-1}(\mathcal{E}_j) + G^{-1}(\mathcal{E}_* \vee \varepsilon)}\right)^{\tilde{m}-2}$$

for all j and for all $\tilde{m} \geq 2$. Thus we can apply Lemma 2.1 to obtain the moment bound

$$\|\mathbf{Z}\|_m \leq \left\|\sup_j \frac{|(P_n - P)(\gamma_j - \gamma_*)|}{G^{-1}(\mathcal{E}_j) + G^{-1}(\mathcal{E}_* \vee \varepsilon)}\right\|_m \leq \sqrt{\Delta} + \frac{K\Delta}{G^{-1}(\mathcal{E}_* \vee \varepsilon)} \ .$$

Altogether then

$$(1-\delta)\mathbf{E}\hat{\mathcal{E}} \leq 2\delta H\left(\sqrt{\frac{\Delta}{\delta^2}} + \frac{K\Delta}{\delta G^{-1}(\mathcal{E}_* \vee \varepsilon)}\right) + (1+\delta)(\mathcal{E}_* \vee \varepsilon) \ .$$

□

Proof of Theorem 4.1. (i) In the power tail case, we define

$$\mathcal{E}_*^\tau := \mathcal{E}_* \vee \tau \ ,$$

where τ is a strictly positive number, and

$$\mathbf{Z} := \frac{|P_n\left((\hat{\gamma}^c - \gamma_*^c)\, 1\left\{\Gamma \leq K\right\}\right)^c|}{C\left(\hat{\mathcal{E}}^{\frac{1}{2\kappa}} + (\mathcal{E}_*^\tau)^{\frac{1}{2\kappa}}\right)} \ .$$

Then we have

$$\begin{aligned}
\hat{\mathcal{E}} &\le |(P_n - P)(\hat{\gamma} - \gamma_*)| + \mathcal{E}_* \\
&= |P_n\left(\hat{\gamma}^c - \gamma_*^c\right)| + \mathcal{E}_* \\
&\le |P_n\left(\left(\hat{\gamma}^c - \gamma_*^c\right)1\left\{\Gamma \le K\right\}\right)^c| + |P\left((\hat{\gamma}^c - \gamma_*^c)1\left\{\Gamma \le K\right\}\right)| \\
&\quad + |P_n\left(\left(\hat{\gamma}^c - \gamma_*^c\right)1\left\{\Gamma > K\right\}\right)| + \mathcal{E}_* \\
&\le C\mathbf{Z}\left(\hat{\mathcal{E}}^{\frac{1}{2\kappa}} + \left(\mathcal{E}_*^\tau\right)^{\frac{1}{2\kappa}}\right) + \mathcal{E}_* + \left(P_n + P\right)\left(\Gamma 1\left\{\Gamma > K\right\}\right) \\
&\le C\mathbf{Z}\left(\hat{\mathcal{E}}^{\frac{1}{2\kappa}} + \left(\mathcal{E}_*^\tau + \left(P_n + P\right)\left(\Gamma 1\left\{\Gamma > K\right\}\right)\right)^{\frac{1}{2\kappa}}\right) \\
&\quad + \mathcal{E}_*^\tau + \left(P_n + P\right)\left(\Gamma 1\left\{\Gamma > K\right\}\right) \ .
\end{aligned}$$

Using Lemma 6.5, we obtain the inequality

$$\begin{aligned}
\hat{\mathcal{E}}^{\frac{1}{2\kappa}} &\le \left(1 + (2\kappa - 1)^{\frac{1}{2\kappa-1}}\right)\left(\frac{C\mathbf{Z}}{2\kappa}\right)^{\frac{1}{2\kappa-1}} \\
&\quad + \left(\mathcal{E}_*^\tau + \left(P_n + P\right)\left(\Gamma 1\left\{\Gamma > K\right\}\right)\right)^{\frac{1}{2\kappa}} \\
&\le \left(1 + (2\kappa - 1)^{\frac{1}{2\kappa-1}}\right)\left(\frac{C\mathbf{Z}}{2\kappa}\right)^{\frac{1}{2\kappa-1}} \\
&\quad + \left(\mathcal{E}_*^\tau\right)^{\frac{1}{2\kappa}} + \left(\left(P_n + P\right)\left(\Gamma 1\left\{\Gamma > K\right\}\right)\right)^{\frac{1}{2\kappa}} \ ,
\end{aligned}$$

where for the second step we used the elementary observation $a^{2\kappa} + b^{2\kappa} \le (a+b)^{2\kappa}$ for $a, b \ge 0$, $\kappa > 1/2$. Now we will first compute the moments of $\mathbf{Z}$ by an application of Bernstein's inequality. We know that

$$P\left|\left(\gamma_j^c - \gamma_*^c\right)1\left\{\Gamma \le K\right\}\right|^m \quad \le \quad K^{m-2}P\left[\left(\left(\gamma_j^c - \gamma_*^c\right)1\left\{\Gamma \le K\right\}\right)^2\right]$$

and

$$\begin{aligned}
P\left[\left(\left(\gamma_j^c - \gamma_*^c\right)1\left\{\Gamma \le K\right\}\right)^2\right] &= P\left[\left(\gamma_j^c - \gamma_*^c\right)^2 1\left\{\Gamma \le K\right\}\right] \\
&\le P\left[\left(\gamma_j^c - \gamma_*^c\right)^2\right] \\
&= \sigma^2\left(\gamma_j - \gamma_*\right) \\
&= \sigma^2\left(\left(\gamma_j - \gamma_0\right) - \left(\gamma_* - \gamma_0\right)\right) \\
&\le \left(\sigma\left(\gamma_j - \gamma_0\right) + \sigma\left(\gamma_* - \gamma_0\right)\right)^2 \ ,
\end{aligned}$$

which by the margin condition

$$\begin{aligned}
&\le \left(C \cdot \left(P\left(\gamma_j - \gamma_0\right)\right)^{1/2\kappa} + C \cdot \left(P\left(\gamma_* - \gamma_0\right)\right)^{1/2\kappa}\right)^2 \\
&= C^2 \cdot \left(\mathcal{E}_j^{1/2\kappa} + \mathcal{E}_*^{1/2\kappa}\right)^2 \ .
\end{aligned}$$

Thus for all j,

$$\begin{aligned}
P\left|\frac{\left(\gamma_j^c-\gamma_*^c\right)1\left\{\Gamma\leq K\right\}}{C\left(\mathcal{E}_j^{1/2\kappa}+\left(\mathcal{E}_*^\tau\right)^{1/2\kappa}\right)}\right|^m &\leq \left(\frac{K}{C\left(\mathcal{E}_j^{1/2\kappa}+\left(\mathcal{E}_*^\tau\right)^{1/2\kappa}\right)}\right)^{m-2}\\
&\leq \left(\frac{K}{C\left(\mathcal{E}_*^\tau\right)^{1/2\kappa}}\right)^{m-2}\\
\Rightarrow P\left|\frac{\left(\left(\gamma_j^c-\gamma_*^c\right)1\left\{\Gamma\leq K\right\}\right)^c}{C\left(\mathcal{E}_j^{1/2\kappa}+\left(\mathcal{E}_*^\tau\right)^{1/2\kappa}\right)}\right|^m &\leq 2\cdot\left(\frac{2K}{C\left(\mathcal{E}_*^\tau\right)^{1/2\kappa}}\right)^{m-2} ,
\end{aligned}$$

and we can apply Lemma 2.1 for loss functions

$$\left(\left(\gamma_j^c-\gamma_*^c\right)1\left\{\Gamma\leq K\right\}\right)^c$$

and parameter distances

$$d(f_j,f_*):=C\left(\mathcal{E}_j^{1/2\kappa}+\left(\mathcal{E}_*^\tau\right)^{1/2\kappa}\right)\leq C\cdot\left(\mathcal{E}_*^\tau\right)^{1/2\kappa}$$

to obtain

$$\|\mathbf{Z}\|_m=\left\|\frac{P_n\left(\left(\hat{\gamma}^c-\gamma_*^c\right)1\left\{\Gamma\leq K\right\}\right)^c}{C\left(\hat{\mathcal{E}}^{1/2\kappa}+\left(\mathcal{E}_*^\tau\right)^{1/2\kappa}\right)}\right\|_m \leq 2\left(\sqrt{\Delta}+\frac{K\Delta}{C\left(\mathcal{E}_*^\tau\right)^{1/2\kappa}}\right) .$$

Now to compute the moments of

$$\left(P_n+P\right)\left(\Gamma 1\left\{\Gamma>K\right\}\right)^{\frac{1}{2\kappa}} ,$$

we proceed as follows for $m\geq 2\kappa$ (using that $\kappa\geq 1/2$):

$$\begin{aligned}
&\left\|\left(\left(P_n+P\right)\left(\Gamma 1\left\{\Gamma>K\right\}\right)\right)^{1/2\kappa}\right\|_m\\
=\ &\left(\mathbf{E}\left[\left(\left(P_n+P\right)\left(\Gamma 1\left\{\Gamma>K\right\}\right)\right)^{m/2\kappa}\right]\right)^{1/m}\\
\leq\ &\left(2^{m/2\kappa-1}\mathbf{E}\left[\left(P_n+P\right)\left(\Gamma^{m/2\kappa}1\left\{\Gamma>K\right\}\right)\right]\right)^{1/m}\\
=\ &2^{1/2\kappa}\left(P\left(\Gamma^{m/2\kappa}1\left\{\Gamma>K\right\}\right)\right)^{1/m} .
\end{aligned}$$

By Lemma 6.2, for $m<2s\kappa$, this has an upper bound in

$$2^{1/2\kappa}\left(\frac{2s\kappa}{2s\kappa-m}\right)^{1/m}M^{s/m}K^{1/2\kappa-s/m} .$$

Thus we find that for $m\in[2\kappa,\min\{1+\log(p),2s\kappa\})$,

$$\left\|\left(\hat{\mathcal{E}}\right)^{\frac{1}{2\kappa}}\right\|_m\leq\left(\mathcal{E}_*^\tau\right)^{\frac{1}{2\kappa}}+A(\kappa)\cdot C^{\frac{1}{2\kappa-1}}\cdot\left(\sqrt{\Delta}+\frac{K\Delta}{C(\mathcal{E}_*^\tau)^{\frac{1}{2\kappa}}}\right)^{\frac{1}{2\kappa-1}}$$

$$+B(\kappa, s, m) \cdot M^{s/m} K^{1/2\kappa - s/m} ,$$

where

$$A(\kappa) := \frac{1 + (2\kappa - 1)^{\frac{1}{2\kappa-1}}}{\kappa^{\frac{1}{2\kappa-1}}} ,$$

$$B(\kappa, s, m) := 2^{1/2\kappa} \left(\frac{2s\kappa}{2s\kappa - m} \right)^{\frac{1}{m}} .$$

If we now apply the straightforward bound

$$\left(\sqrt{\Delta} + \frac{K\Delta}{C(\mathcal{E}_*^\tau)^{\frac{1}{2\kappa}}} \right)^{\frac{1}{2\kappa-1}} \leq \left(\sqrt{\Delta}\right)^{\frac{1}{2\kappa-1}} + \left(\frac{K\Delta}{C(\mathcal{E}_*^\tau)^{\frac{1}{2\kappa}}} \right)^{\frac{1}{2\kappa-1}}$$

and minimize the upper bound over $K \geq 0$ (using Lemma 6.3), we obtain the desired oracle inequality for the power tail case.

(ii) If we assume the exponential moment condition instead of power tails, we can take

$$\mathbf{Z} := \frac{|P_n\left((\hat{\gamma}^c - \gamma_*^c)\right)|}{C\left(\hat{\mathcal{E}}^{\frac{1}{2\kappa}} + (\mathcal{E}_*^\tau)^{\frac{1}{2\kappa}}\right)}$$

and we obtain the same bound for $\|\mathbf{Z}\|_m$ as before, but no term stemming from $\Gamma 1\{\Gamma > K\}$. This yields the desired risk moment inequality. The corresponding risk tail bound also comes straight from applying Bernstein's inequality (13) to $\mathbf{Z}$.

□

6.4. Proofs for Section 5

Proof of Lemma 5.1. Regard the term

$$\frac{\sqrt{a} + \sqrt{b}}{\sqrt{\frac{a+b}{2}} + \sqrt{b}}$$

for $a, b \geq 0$. Some simple calculus shows that for fixed a, this ratio attains its maximum for $b = 0$; thus

$$\frac{\sqrt{a} + \sqrt{b}}{\sqrt{\frac{a+b}{2}} + \sqrt{b}} \leq \sqrt{2}$$

for all $a, b \geq 0$. Using this inequality, and the definition of the Hellinger distance, we can now compute

$$\begin{aligned} h^2(\bar{f}, f) &= \frac{1}{2}\int \left(\sqrt{\frac{f+f_*}{2}} - \sqrt{f}\right)^2 d\mu \\ &= \frac{1}{8}\int \left(\sqrt{f_*} - \sqrt{f}\right)^2 \cdot \frac{\left(\sqrt{f}+\sqrt{f_*}\right)^2}{\left(\sqrt{\frac{f+f_*}{2}} + \sqrt{f}\right)^2} d\mu \\ &\leq \frac{1}{4}\int \left(\sqrt{f_*} - \sqrt{f}\right)^2 d\mu \\ &= \frac{1}{2}h^2(f, f_*) \ . \end{aligned}$$

The triangle inequality now gives us

$$\begin{aligned} h(f, f_0) &\leq h(f, \bar{f}) + h(\bar{f}, f_0) \\ &\leq \frac{1}{\sqrt{2}}h(f, f_*) + h(\bar{f}, f_0) \\ &\leq \frac{1}{\sqrt{2}}h(f, f_0) + \frac{1}{\sqrt{2}}h(f_*, f_0) + h(\bar{f}, f_0) \\ \Rightarrow \left(1 - \frac{1}{\sqrt{2}}\right) h(f, f_0) &\leq \frac{1}{\sqrt{2}}h(f_*, f_0) + h(\bar{f}, f_0) \ , \end{aligned}$$

from which the statement of the lemma follows. □

Proof of Lemma 5.2. Clearly, we have $f_* = f_p = f_0$, and $\mathcal{E}_* = 0$.

The margin condition holds with $\kappa = 1$, $C^2 = 4\sigma_\epsilon^2$, $\sigma_\epsilon^2 := E\epsilon^2 = 2/((s-2)(s-1))$, and $d(f, f_0) \geq \sigma^2(\gamma_f - \gamma_{f_0})$. When $p \geq \sqrt{n}$, moreover, the tail condition holds with $M = 2$, since

$$\Gamma(Z_i) = \max_{1 \leq j \leq p} |\gamma_j^c(Z_i) - \gamma_*^c(Z_i)| = \max_{1 \leq j \leq p} 2|\epsilon_i f_j(X_i)| = 2|\epsilon_i| n^{\frac{1}{2s}} \{1 \leq i \leq p-1\}$$

and thus

$$P(\{\Gamma > K\}) = \frac{1}{p-1} P(2|\epsilon| n^{\frac{1}{2s}} > K) = \frac{1}{p-1}\left(\frac{1}{1 + K/(2n^{\frac{1}{2s}})}\right)^s \leq 2^s K^{-s} \ .$$

We also have for all $u > 0$, and $n \geq 2^{2s}$,

$$\begin{aligned} \mathbf{P}(\max_{1 \leq j \leq p-1} 2\epsilon_j \leq (1+u)n^{\frac{1}{2s}}) &= \left(1 - \frac{1}{2}\left(\frac{1}{1 + (1+u)n^{\frac{1}{2s}}/2}\right)^s\right)^{p-1} \\ &= \left(1 - \frac{1}{2}\left(\frac{1}{n^{\frac{1}{2s}}(n^{-\frac{1}{2s}} + (1+u)/2)}\right)^s\right)^{p-1} \end{aligned}$$

$$\leq \left(1 - \frac{1}{2}\left(\frac{1}{n^{\frac{1}{2s}}(1+u/2)}\right)^s\right)^{p-1} \leq \exp[-2^{-1}(1+u/2)^{-s} \cdot (p-1)/\sqrt{n}].$$

It follows that with probability at least $1 - \exp[-2^{-1}(1+u/2)^{-s} \cdot (p-1)/\sqrt{n}]$,

$$\min_{1 \leq j \leq p} P_n(\gamma_j) < P_n(\gamma_0) - un^{\frac{1}{2s}}.$$

Thus with probability at least $1 - \exp[-2^{-1} \cdot (p-1)/\sqrt{n}]$, we have that $\hat{\gamma} \neq \gamma_0$.

But for $\gamma_j \neq \gamma_0$,

$$P(\gamma_j - \gamma_0) = P|f_j|^2 = \frac{1}{n}\sum_{i=1}^n f_j^2(X_i) = \frac{1}{n} f_j^2(X_j) = n^{-\frac{s-1}{s}}.$$

□

Proof of Lemma 5.3. Like in the proof of Theorem 4.1, define

$$\mathbf{Z} := \frac{|(P_n - P)(\hat{\gamma} - \gamma_*)|}{C(\sqrt{\hat{\mathcal{E}}})},$$

whenever $\hat{\mathcal{E}} > 0$. Then

$$\sqrt{\hat{\mathcal{E}}} \leq \sqrt{\mathcal{E}_*} + C\mathbf{Z}.$$

For any n constants $(b_1, \ldots, b_n)$, we know that

$$\left\| \frac{1}{n}\sum_{i=1}^n b_i \epsilon_i \right\|_s \leq \frac{c_s M}{n} \sqrt{\sum_{i=1}^n b_j^2}$$

(see Whittle [26] or Appendix A of van der Vaart and Wellner [23]). Hence

$$\|\mathbf{Z}\|_s \leq c_s p^{1/s} M/\sqrt{n},$$

and thus

$$\left\| \sqrt{\hat{\mathcal{E}}} \right\|_s \leq \sqrt{\mathcal{E}_*} + Cc_s p^{1/s} M/\sqrt{n}.$$

□

Acknowledgements

This work was partly supported by the Swiss National Science Foundation (SNF Grant 20PA21-120050/1). We are also grateful to an anonymous referee whose detailled suggestions considerably improved the presentation of this paper.

References

[1] Audibert, J.-Y. (2003). Aggregated estimators and empirical complexity for least square regression. Preprint no. 805, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7. MR2096215

[2] Audibert, J.-Y. (2006). A randomized online learning algorithm for better variance control. In *Proceedings of the 19th Annual Conference on Learning Theory*, pp. 392–407. MR2280620

[3] Audibert, J.-Y. (2007). Progressive mixture rules are deviation suboptimal. *Advances in Neural Information Processing Systems*.

[4] Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Prob. Theory and Rel. Fields* **113**, 3: 301–413. MR1679028

[5] Bartlett, P.L. and Mendelson, S. (2006). Empirical minimization. *Prob. Theory and Rel. Fields* **135**, 3: 311–334. MR2240689

[6] Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. *Festschrift for Lucien Le Cam.* Springer, New York. MR1462939

[7] Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *C.R. Acad. Sci. Paris* **334**, 6: 495–500. MR1890640

[8] Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35**, 4: 1674–1697. MR2351101

[9] Chesneau, C. and Lecué, G. (2006). Adapting to Unknown Smoothness by Aggregation of Thresholded Wavelet Estimators. ArXiv preprint math.ST/0612546.

[10] Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A Distribution-free Theory of Nonparametric Regression.* Springer, New York. MR1920390

[11] Györfi, L., and Wegkamp, M. (2008). Quantization for Nonparametric Regression. *IEEE Trans. Inform. Theory* **54**, 2: 867–874. MR2444565

[12] Juditsky, A., Rigollet, P., and Tsybakov, A. (2008). Learning by mirror averaging. *Ann. Statist.* **36**, 5: 2183–2206. MR2458184

[13] Juditsky, A. B., Nazin, A. V., Tsybakov, A. B., and Vayatis, N. (2005). Recursive aggregation of estimators by the mirror descent method with averaging. *Problemy Peredachi Informatsii* **41**, 4: 78–96. MR2198228

[14] Koltchinskii, V. Local Rademacher Complexities and Oracle Inequalities in Risk Minimization. 2004 IMS Medallion Lecture, July 2005.

[15] Lecué, G. (2007). Suboptimality of Penalized Empirical Risk Minimization in Classification. *Proceedings of the 20th Annual Conference On Learning Theory. Lecture Notes in Artificial Intelligence 4539*, 142–156. Springer, Heidelberg. MR2397584

[16] Lecué, G. (2007). Optimal rates of aggregation in classification under low noise assumption. *Bernoulli* **13**, 4: 1000–1022. MR2364224

[17] Lee, W.S., Bartlett, P.L. and Williamson, R.C. (1998). The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory* **44**, 5: 1974–1980. MR1664079

[18] Mendelson, S. (2007). Obtaining fast error rates in nonconvex situations. *J. Complexity* **24**: 380–397. MR2426759

[19] Mendelson, S.. Lower bounds for the empirical minimization algorithm. To appear in IEEE Transactions on Information Theory. MR2451042

[20] Rigollet, P. (2006). Inégalités d'oracle, agrégation et adaptation. Ph.D. thesis, Université Paris-VI.

[21] Tsybakov, A. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32**, 1: 135–166. MR2051002

[22] van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

[23] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*, Springer, New York. MR1385671

[24] van der Vaart, A. W., Dudoit, S. and van der Laan, M. J. (2006). Oracle inequalities for multi-fold cross validation. *Statistics & Decisions* **24**: 351–371. MR2305112

[25] Vapnik, V. and Chervonenkis, A. (1974). *Theory of Pattern Recognition.* Nauka, Moscow (in Russian). MR0474638

[26] P. Whittle (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability and its Applications* **5**, 3: 302–305. MR0133849